基于回归分析的网络恐怖信息主题爬虫*

黄炜[1] 张展程[1] 朱彬[1] 李岳峰[1] 陆薇[2]
[1] 湖北工业大学经济与管理学院 武汉 430068 [2] 国家电网武汉市东湖新技术开发区供电公司 武汉 430073

摘要：[目的/意义]针对目前从开源网络信息中采集网络恐怖信息难、采集效率低的问题，提出一种回归分析法，以综合语义相关与网页重要性两个因素，从而提高网络恐怖信息的采集效率。[方法/过程]通过分析、比较主题爬虫的特性，结合网络恐怖信息的特点，找出 PageRank 算法和 TF-IDF 算法中适用于恐怖信息采集的优点，并结合回归分析法，将恐怖信息的采集策略进行相关度预测，用预测结果反馈调节信息的采集过程。[结果/结论]网络恐怖信息采集要兼顾采集的数量和质量，在传统主题爬虫算法的基础上进行改进，提出针对于开源网络恐怖信息采集的爬虫优化算法，可以提高信息采集效率。

关键词：主题爬虫 回归分析 网络反恐 语义相似度
分类号：G206
DOI：10.13266/j.issn.0252-3116.2018.04.016

1 引言

网络空间与恐怖主义相结合的概念最早由美国加州情报与安全资深研究员 B. Coins 于 1997 提出，他认为网络恐怖主义是网络与恐怖主义相结合的产物[1]。如今大数据时代背景下，人们主要的信息来源是网络，网络恐怖主义也利用互联网时代的便捷性不断发展。恐怖分子们通过互联网来发布一些网络恐怖信息，一方面造成社会恐慌，另一方面通过互联网发布恐怖活动的组织策划信息，为一场恐怖活动的发生提供先决的信息条件。李本先等认为从互联网方面击退恐怖主义是网络反恐研究的重要方向[2]。李欧认为网络反恐的特性——网络技术的脆弱性，网络活动的隐蔽性，网络资源的丰富性和开放性，网络普及性和广泛性，让恐怖分子得以利用，成为恐怖分子开展恐怖活动的"第二战场"[3]。

随着信息化时代的快速发展，当前网络恐怖主义呈现全球化，国际性恐怖主义组织与地区性恐怖主义组织紧密相连[4]。在 2016 年乌镇举行的第三届世界互联网大会的"网络反恐论坛"上就提出加强国际间的合作、联手打击网络恐怖主义[5]。虽然网络反恐得到世界各国的重视，但恐怖组织战术越来越灵活多变，行动能力越来越专业化，组织越来越信息化[6]。面对新特点，网络安全工作者和反恐相关部门想在茫茫的数据海洋中搜集恐怖信息数据的需求愈加强烈。网络反恐数据的收集是网络反恐的基础，定制特异性的垂直搜索引擎技术是网络恐怖信息采集的关键[7]。在对反恐领域的相关文献研究的基础上，本文从定性和定量两个角度，将网络恐怖信息特征总结为数字化特征和本质特征两大类，如图 1 所示，并在原有的主题爬虫算法上进行改良，提出针对网络反恐数据采集的特异性主题爬虫。

恐怖信息的纵向特点是用来量化处理和判别网络恐怖信息可信度的要素，称之为数字化特征，可以用算法量化计算，横向是用来形容网络恐怖信息的抽象特征，只能根据经验人为的进行判定，称之为本质特征。本文针对网络恐怖信息的数字化特征提出回归分析模型将各个恐怖相关因素逐渐适应到一条曲线上，综合判断网络恐怖信息的各项特征，从而提高主题爬虫对网络恐怖信息的采集精度。

* 本文系国家自然科学基金项目"微博环境下实时主动感知网络舆情事件的多核方法研究"（项目编号：71303075）和"大数据环境下基于特征本体学习的无监督文本分类方法研究"（项目编号：71571064）研究成果之一。
作者简介：黄炜（ORCID：0000-0002-5804-9371），教授，博士，硕士生导师，E-mail：tonny hw@163.com；张展程（ORCID：0000-0002-7533-4764），硕士研究生；朱彬（ORCID：0000-0002-1073-0379），本科生；李岳峰（ORCID：0000-0001-5173-9575），教授，博士；陆薇（ORCID：0000-0001-6270-8846），助理工程师。
收稿日期：2017-08-21 修回日期：2017-11-16 本文起止页码：121-129 本文责任编辑：王善军

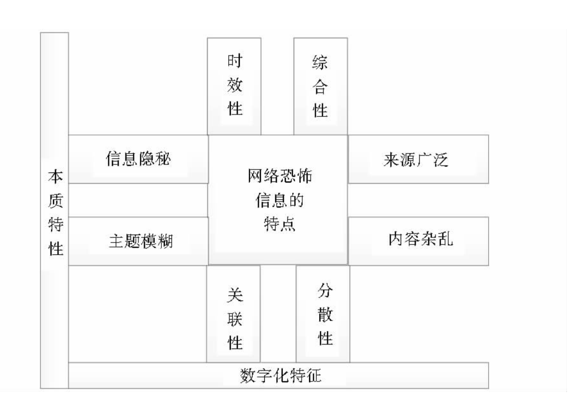


图 1 网络恐怖信息的纵横图

2 主题爬虫与算法回顾

主题爬虫(FocusedCrawler)是专门用来搜集互联网上具有特定主题文档信息的智能主体,能自动地在互联网上搜索爬行,并将搜集的主题信息返回给服务器[8]。主题爬虫的工作流程图如图 2 所示:

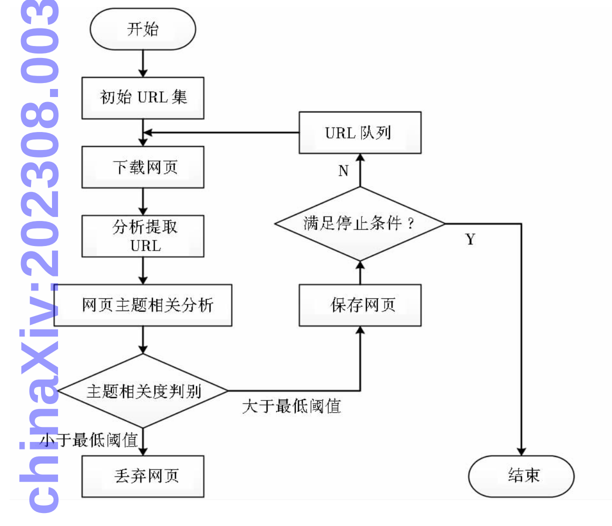


图 2 主题爬虫的工作流程

主题爬虫采集网络信息的关键在于如何精确地分析网页内容与主题之间的相关程度,以及使用怎样的采集策略才能使采集过程更加高效及时。主题爬虫算法研究分为两类:一类是基于链接结构[9],一类是基于内容评价[10]。前者的代表算法是 PageRank 算法,后者的代表有 TF-IDF 算法。

基于链接结构的网页评价算法中,代表的是 PageRank 算法,其中杨彬等提出基于概念的权重 PageRank 改进算法[11],林泓等基于用户点击网页内的各个链接的概率不等的情况下提出对 PageRank 改进算法,从而避免主题漂移现象[12],何明等提出语义相似的 PageRank 改进算法[13],王钟斐等提出基于锚文本相似度的 PageRank 改进算法[14],王建雄在传统 PageRank 算法的基础上进行改进,通过计算超链接与领域向量的相似度来抑制主题漂移,引入时间因子以及站内外区分因子来提高与主题相关 URL 的权重,从而提高信息采集效率[15],王冲等提出基于用户兴趣与主题相关的 PageRank 改进算法[16]。PageRank 算法是主题爬虫算法的核心,被应用于用户特定主题的垂直搜索引擎中,能够较好地决定采集策略,但是 PageRank 算法以及目前对其改进的算法中,都是根据用户访问的特定主题信息的网络信息[17],基于此对网络中的链接的重要性进行调整,如杨彬和林泓从用户浏览网页的角度,对网页中 URL 队列的权重进行调整达到优化爬行策略的目的,J. M. Maestre 等将 PageRank 应用于对目标群体的联合控制中[18],但网络恐怖信息页面并不是大众用户频繁访问的页面,因此不能简单通过链接的点击量和访问量及链入、链出网页的数量来确定网页的重要性,甚至网页链接数量相对多的网页与网络恐怖无关的几率更大,因为网络恐怖信息的特性就是隐秘性强,它不同于传统搜索引擎搜索的并按照受欢迎度进行排序的网页,所以网页受欢迎度并不能作为评判网页恐怖主题的关键指标。

在网络主题爬虫搜索策略的算法研究中,除了上述的基于 PageRank 改进的链接评价外,主题爬虫算法中另一类是对网页中内容进行评价,代表算法有 TF-IDF 算法。路永和等将 TW 与 TF-IDF 结合作为新的特征权重算法[19],王景中等将正则表达式和语义分析技术相结合,从而实现对 TF-IDF 算法的改进[20]。在对网页内容与主题进行评价上,改进的 TF-IDF 充分利用原网页中的标签、描文本、过滤虚词的方式来充分调整对网页中关键词赋予权重的值,从而比较精确地判断网页与主题的相关性。但网络恐怖信息的特点是主题模糊,内容杂乱和分散,信息的不确定性较大,目前改进的 TF-IDF 算法不能很好地利用网页链接之间的相互关系,从而不能很好地对恐怖主题相关的网页做一个关联性判定,不能按照网络恐怖信息的关联性进行采集。

3 系统框架设计

面对传统主题爬虫算法所存在的偏激问题,本文结合改进的 PageRank 算法和改进的 TF-IDF 算法提出回归分析预测模型。由于判断网络信息是否与恐怖相关不能仅由一个或几个主题关键词来判别,更不能只根据网页链接数和点击量来决定,而是要通过一定的分析将这些相关因素有机结合起来才能够提高网络恐

怖信息采集的质量和效率,本文吸收了基于本体的网络群体性事件主题发现模型的方法[21],建立自己的网络本体恐怖信息库,从而充分利用已经采集的网络恐怖信息数据,为回归性分析网络恐怖信息提供数据依据。

传统的主题爬虫系统需要事先提供一定数量的初始URL和主题词,初始URL是为解析得到其它的网页页面,并提取网页中二级乃至多级的URL从而形成初始URL队列,循环上述过程得到待爬取的URL库,而主题词用于判断爬取的文本与主题的相关性,但是选取一些质量高、能够比较全面地描述网络恐怖信息的主题关键词就显得十分困难。本文提出使用回归分析预测的方法来解决采集网络恐怖信息的困境,其具体思路如下:①通过给定若干网络恐怖信息页面,由回归分析预测模块分析筛选出恐怖主题关键词,并将关键词和页面链接数量存储起来,建立恐怖信息词表,词表中的关键词和链接数为该条恐怖关键词的标签。②通过恐怖关键词与涉恐URL回归,设置回归分析偏移量来控制每次网络恐怖信息采集关键词的数量,为主题爬虫提供关键词和初始URL队列。③由初始URL解析到对应的网页,由采集模块采集网页的内容,由分析模块分析出网页的主题关键词的个数和URL链接及其它网页信息。④通过恐怖网页信息相似性回归,将采集的网页按照一定规则进行解析,把恐怖信息相关的因素通过回归函数回归到一条曲线上,根据与权威恐怖信息网页的拟合优度的高低来选择网络恐怖信息采集的优先次序。⑤通过回归分析把拟合优度高的网页中的关键词和网页链接数以及其它因素存储到网络本体恐怖信息库中,拟合优度较低的但在设定最低阈值内的恐怖信息网页经过人工审核的方式添加到网络恐怖信息库中,为下次回归分析采集提供参考。回归分析主题爬虫结构图如图3所示:

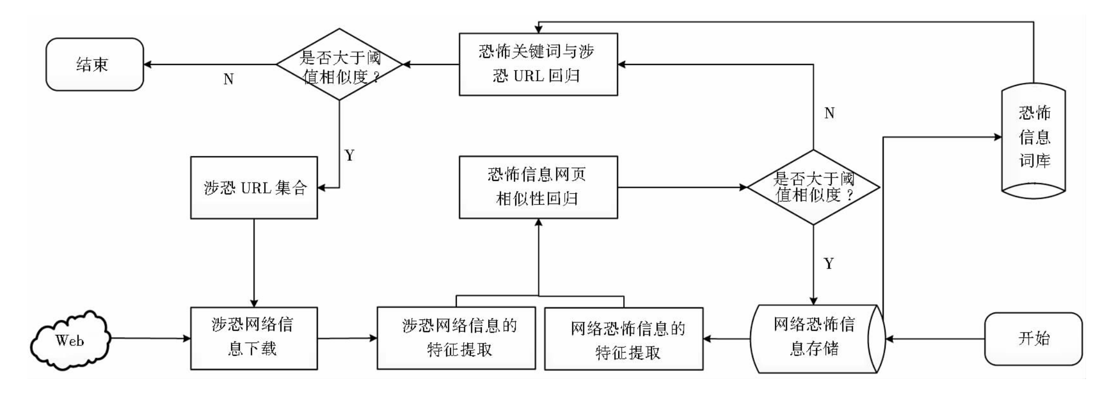


图3 涉恐信息主题爬虫结构

4 回归分析

4.1 改进PageRank算法

网络反恐信息采集过程中可以量化的有关联性和综合性两大特征,而这两大特征可以采用改进的PageRank算法来度量。PageRank算法作为Google公司对于网页评价的重要算法,其最显著的特点就在于在Web链接方面十分客观地反映网页的重要程度[22],原理在于其网页通过链接"投票",得出网页入链接与出链接的情况,衡量其"链接流行度",即对于某个网页节点来说,链入的网页数量越大,说明此网页重要性越大。网页重要程度通过PR值量化表现,PR值区间为[0,10],PR值越大表明其越受欢迎[23]。

对于某个涉恐网页M来说,PageRank算法对它遵循两个假设:

(1)数量假设:如果涉恐网页M链入的恐怖网页的链接越多及关联性强,则表示其影响大,重要性高。

(2)质量假设:网络恐怖信息越重要的网页给予的恐怖权重高,如果有其它网页对其进行链接,其重要程度相对来说越高。

设有网页 $I_1, I_2, I_3, \cdots, I_n$ 对网页M有链入,第i个网页的总链接数为L(i),根据网页对链接的权重分配来计算,网页M的PR值如式(1)所示:

$$PR(M) = \frac{PR(I_1)}{L(I_1)} + \frac{PR(I_2)}{L(I_2)} + \frac{PR(I_3)}{L(I_3)} + \cdots + \frac{PR(I_n)}{L(I_n)} \qquad 式(1)$$

同时由于存在不链入任何其他链接的网页,也叫做"死链接",使得公式出现错误,故添加阻尼系数(damping factor)d对公式进行修正,d一般取值0.85[24]。随机冲浪模型也证实阻尼系数的用处,表示一个网页漫游者不会一直点击一个链接,而会随机的跳转到其他网页,保证网页对于每个链出的链接权重分配平均。修改后如式(2)所示:

$$PR(M) = (1-d) + d\left(\frac{PR(I_1)}{L(I_1)} + \frac{PR(I_2)}{L(I_2)} + \frac{PR(I_3)}{L(I_3)} + \cdots + \frac{PR(I_n)}{L(I_n)}\right) \quad 式(2)$$

4.2 改进 TF-IDF

语义分析是主题相关度分析的重要因素,单独对网页重要性分析时不会考虑主题词是否与网页内容相匹配,故可能出现“主题漂移”现象。本文采用 TF-IDF 这种语义分析算法通过对网页文本内容进行特征词分析,通过网络恐怖特征词的分布情况对网页进行主题相关度计算。TF-IDF 算法在创始之初存在着一定的缺陷,Z. H. Deng 等[25]提出的替代的 CRF(category relevance factors),赵小华[26]等提出的运用特征选择修正函数权重的 TF-IDF-CHI 算法均对初始算法进行一定的改进。

在此基础上王景中等[20]提出的改进 TF_IDF 算法中指出对于特征词内的关键字赋予权重是非常重要的,因为在特征词中作出贡献的关键字可能存在“的”“呢”等虚词,使得权重赋予无意义,故需要剔除虚词,对关键字与非关键字进行不同的权重赋予。在锚文本、title 和 meta 等标签中特征词占比例很高[27],考虑到上述标签单一出现贡献度会出现转移,采用评价加权和累积计算的方法得出权重公式如式(3)所示:

$$T_{wf}(k) = \frac{\sum_{i=0}^{n} m(i)}{\sum_{j=1}^{N} m(j)} \quad 式(3)$$

$m(i)$是第 i 个标签的权重数值,$T_{wf}(k)$指第 k 个词的平均累加权重数值,$\sum_{i=0}^{n} m(i)$表示第 k 个词在所在标签的累加权重,$\sum_{j=1}^{N} m(j)$表示在整个页面包含的上述标签的权值总和。根据以往对网络恐怖信息的研究得到比较合理的标签权值函数 $m(i)$ 如式(4)所示[27]:

$$m(i) = \begin{cases} 10, & \text{title} \\ 8, & \text{meta} \\ 6, & \text{H, a} \\ 3, & \text{其他} \end{cases} \quad 式(4)$$

路永和等[19]针对特征词在文本类别内与类别之间的分布情况,提出对于特征词与文档类别,不包含该特征词但是属于该类的文档数 C 与包含该特征词但是不属于该类的文档数 B,加上常用的特征选择评估函数卡方值 CHI,总结出 C 越小,B 越小,说明类内分布越均匀分散,而类间分布情况高度聚集,那么此特征项的特征权重数值就越大。得出这种权重如式(5)所示:

$$TW_{(i)} = \log CHI_i \times \frac{1}{\log(B_i \times C_i)} \quad 式(5)$$

i 为恐怖信息的特征项,CHI_i 为该恐怖特征项的 CHI 值,B_i 为包含 i 但不属于该类的涉恐文本数,C_i 为不包含 i 但属于该类的涉恐文本数。由于网络恐怖信息的模糊性和隐秘性对于某些恐怖信息特征项,$B_i \times C_i$ 可能为 0,故加入一个相对于 $B_i \times C_i$ 较小的常数 λ 得出权重公式如式(6)所示:

$$TW_{(i)} = \log CHI_i \times \frac{1}{|\log(B_i \times C_i + \lambda)|} \quad 式(6)$$

最后基于特征词中的关键字贡献度和文档类别间的分布关系得出如下计算语义相关度的综合 TF-IDF 公式如式(7)所示[16]:

$$W_{ik} = TW_{(t)} \times T_{wf}(k) \times \log tf_{ik} \times idf_{ik} = TW_{(t)} \times T_{wf}(k) \times \log(\varepsilon + N/n_k) \quad 式(7)$$

W_{ik}表示包含第 i 个恐怖特征词的第 k 篇涉恐文档的权重,$TW_{(t)}$表示对于涉恐文档类别间恐怖特征词分布情况的项修正权重,$T_{wf}(k)$表示第 k 篇文档的关键字权重,N 表示文档总数,n_k 表示含有特征词的文档数。由于涉恐网络恐怖信息的文本长度不同,涉恐文本长度长的恐怖关键词的权值会偏大,为了解决这个问题,通过取对数 $\log tf_{ik} \times idf_{ik} = \log(\varepsilon + N/n_k)$ 的方式进行一种标准化,减少因文本过长导致权重偏大所带来的影响。

4.3 恐怖关键词与涉恐 URL 回归

主题爬虫的工作原理是对种子 URL 集合进行解析,并提取对应的 URL 队列,下载 URL 队列对应的页面,解析页面所包含的一系列 URL 并对 URL 集合进行扩展。每次爬虫循环都会形成下一次采集的 URL 集合,因此 URL 对于主题爬虫信息采集的效率及质量有比较大的影响[28]。

回归分析法核心是由初始 URL 扩展到其它 URL 时,其它 URL 所对应的网页内容与主题的相关程度与最初的 URL 所对应的与主题的相关程度线性相关。简单说就是主题相关度越高的网页,其内包含的 Web 链接内含有的的网页内容与主题的相关性也要更高一些。基于这种思想本文先将初始的 URL 进行解析,提取出其中的 URL 和网页中的词汇,并建立索引 URL 的单向索引和关键词的多重索引。剔除掉重复关键词中的虚词等重复率高但与主题明显偏离的词汇,制成主题词汇表,再引入 PageRank 算法中 URL 的重要性 PR(M),将剔除后的关键词表标号,与 URL 重要性 PR(M)建立对应关系如图 4 所示:

黄炜，张展程，朱彬，等.基于回归分析的网络恐怖信息主题爬虫[J].图书情报工作,2018,62(4):121－129.

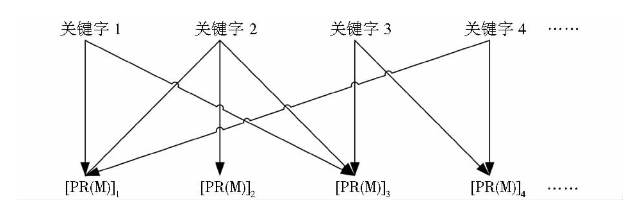


图 4　URL 重要性建立关系图

$$P_1 = \{[PR(M)]_1 + [PR(M)]_3 + [PR(M)]_j + \cdots\} * \rho + c^* W_{ik} \quad 式(8)$$

式(8)表示该次采集过程中每个关键词在众多 URL 中与主题重要程度。P_1 表示的是第一个主题关键字的重要程度，$[PR(M)]_j$ 是关键词在网页 j 中所对应的重要性，W_{ik} 是 TF-IDF 算法中关键词的累计权重，ρ 和 c 是参数，用来调节两个变量之间的关系。

通过上述方法，将 URL 的重要性或者说是受欢迎程度这个因素引入到主题关键词对网页与主题相关性的评估上，就能够对那些在抓取内容中出现次数少但与主题相关度高和在爬取内容中出现次数高但与主题相关度低的关键词赋予了一定的权重，对不同重要性的关键词所对待的程度也不一样，从而提高关键词对网页内容的考核精度。

将网页包含所有关键词的重要程度为记为 X，$X = \sum_{i=1}^{n} p_i$，将网页的重要程度记为 Y，$Y = PR(M)_n$，如式(9)所示：

Y = A + BX + §；

B = ∑xy − n · ∑x∑y/[∑x&sup2 − n · (∑x)&sup2]；

A = ∑y − B · ∑x/n；　　式(9)

进行相关性回归性分析，这就是回归分析 2 模块图。回归分析得到的回归线如图 5 所示：

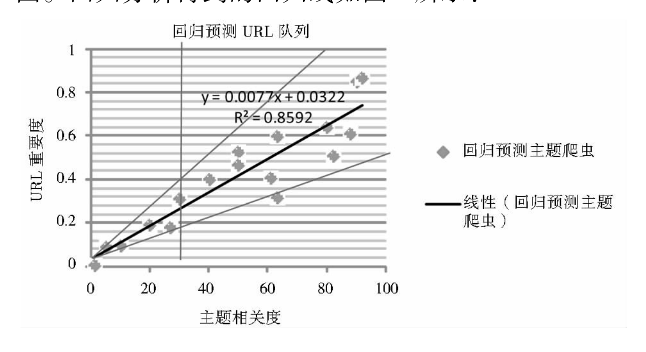


图 5　主题回归预测主题爬虫

再作两条直线分别为式(10)和式(11)所示：

Y1 = (A + v)x + B + §；　　式(10)

Y2 = (A - v)x + B + §；　　式(11)

A 和 B 为待定参数，A 为回归直线的截距；B 代表回归直线的斜率，表现 Y 随 X 每单位变化时的平均变化情况；§ 代表以用户满意度为参考因素的随机误差值。

Q 代表最低主题相关度；V 代表的是容差值。Q 和 V 为主题爬虫事先需要定义的两个阈值用来调节和控制爬取的 URL 范围，Q 是保证页面内容与主题相关度的参数，V 是可允许的最大 URL 与主题的不相关度。爬虫工作时只需爬取 Q 与 Y1 和 Y2 所围成的区域，剔除掉这个区域以外的 URL，将剩下的 URL 作为 URL 集合投入到下一轮的循环中，如图 3 涉恐信息主题爬虫结构图中恐怖关键词与涉恐 URL 回归模块中当 V = 0.25 的时，几乎囊括所有涉恐 URL，通过这样一种方式将会大大减少 URL 的数量，提高 URL 队列的质量，减小了主题爬虫的工作负担，提高系统的运行效率。

4.4　恐怖网页信息相似性回归

网络恐怖信息存在时效性、分散性等特点，依据网络涉恐信息中可能存在的这些潜藏的特性，本文通过建立一条逻辑曲线将同网页中用来判别恐怖信息的各种因素如：关键词、链接数、出现时间、访问量等因素组合起来如公式(12)所示。

$$Z = \beta_0 + \beta_1 \varphi_1 + \beta_2 \varphi_2 + \cdots + \beta_k \varphi_k \quad 式(12)$$

式中 β_i 被称为回归参数，φ_i 是各个恐怖影响因素。回归分析一开始，是人工根据经验或期望对 φ_i 的值进行设定，其值大小如 HITS 算法中的权重一样。随着爬取内容的不断增多，将从网络上爬取下来的涉恐信息进行去重，降噪处理之后，将这一类网络恐怖信息网页对 φ_i 进行打分，按照分数大小对 φ_i 进行调整。

将式(13)通过回归函数如公式(13)所示。

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{e^{-z} + 1} \quad 式(13)$$

函数图如图 6 所示：

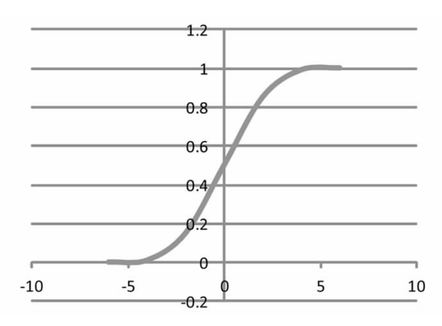


图 6　相似性特征回归

该算法回归模型的特点是变量的范围是从 －∞ 到 ＋∞，但是值域的范围是在(0－1)之间，这样就将多

个恐怖因素转化成一个概率来判断网络恐怖信息的相关性。将($\beta_0, \beta_1, \beta_2, \cdots, \beta_k$)按照赋予了网页链接数的主题关键词的 P_n 来归为一类,每次获取一个新的网页时:①将网页中的词语与链接按照网页模块提取出来。②将每一个网页模块中的 P 值算出,去数据库中匹配到主关键词下对应的 P 值。③将数据库中该 P 值标签下的一组($\beta_0, \beta_1, \beta_2, \cdots, \beta_k$)作为该网页中关键信息的参数,用来计算 Z 值,根据 f(z)所对应的(0 - 1)上的值来判断网页恐怖信息程度。

对于某个网页模块中可能缺少关键信息 φ_k 的这种情况,首先在计算时先将原有的数据库中的 φ_k 补充上去,后再减去缺失这部分恐怖因素与 $\log_v §$(误差值与容错值的对数),以降低因缺失某部分网络恐怖因素对回归分析判别网络恐怖信息可信度程度。最后计算 z 的公式变成公式(15):

$$z = \beta_0 + \beta_1 \varphi_1 + \beta_2 \varphi_2 + \cdots + \beta_k \varphi_k - \frac{\sum_{i=1}^{n} \beta_j \varphi_j}{\log_v §}$$

式(14)

5 实验结果分析与评估

为了验证引入回归分析改进的主题爬虫算法在爬取网络开源恐怖信息上的有效性,实验将改进的语义相似 PageRank 算法,改进的 TF-IDF 算法与回归分析算法进行对比分析来证明引入回归分析算法的主题爬虫采集网络恐怖信息的优越性。

5.1 实验设计

为了验证上述改进算法的优势之处,采集2 000条文本作为实验的数据源,其中 1 000 条文本为是通过网络恐怖信息采集系统采集的,经过部分人工筛选,确定的恐怖信息文本,另外 1 000 条文本为通用爬虫采集的普通网络文本信息,但在每篇普通文本中随机位置写入网络恐怖词库中的恐怖关键词,进而构成恐怖信息文本的实验样本。

5.2 反恐词库建立

反恐词库的建立是回归分析中最重要的一点,将词库中的恐怖主义相关的词汇与链接建立对应关系,才能进一步将新的网页与词库的词汇和标签建立关系,通过词库与标签来开展恐怖信息回归分析,从而判断网页与恐怖主题的相关程度。将爬取的 10 000 条网络恐怖信息数据经过人工分析,建立一张 100 条恐怖主题信息词库表,表的部分如表 1 所示。表中首字母 A、B、C 等是恐怖信息词的类别,词库类别是按照词的属性进行分类,比如地点名词、事件名词、恐怖主义代号、暗语等。字母后面的数字表示的是其序号,序号的位数越多表明这个词库的重要性越高,所对应的权重相对来说也越高。

表 1 网络恐怖信息部分词表

编号	主题词	累计频数	被链接数	标签 p
A001	菲律宾达沃	64	10	菲律宾达沃$^{64}_{10}$
A002	莫赫曼德特区	56	5	莫赫曼德特区$^{56}_{5}$
A003	叙利亚	78	5	叙利亚$^{78}_{5}$
A004	伊德利卜	58	11	伊德利卜$^{58}_{11}$
A005	索马里极端组织	18	13	索马里极端组织$^{18}_{13}$
A0076	迈杜古里	62	7	迈杜古里$^{62}_{7}$
A007	巴基斯坦	24	42	巴基斯坦$^{24}_{42}$
A008	伊拉克	46	58	伊拉克$^{46}_{58}$
A009	伊斯兰国	118	19	伊斯兰国$^{118}_{19}$
A010	圣诞市集	9	13	圣诞市集$^{9}_{13}$
A014	教堂	11	19	教堂$^{11}_{19}$
B001	风暴	22	37	果断风暴$^{22}_{37}$
B002	人质劫持	137	14	人质劫持$^{137}_{14}$
B003	装甲车	214	28	装甲车$^{214}_{28}$
B004	空袭	32	8	空袭$^{32}_{8}$

词表的建立过程如图 7 所示:

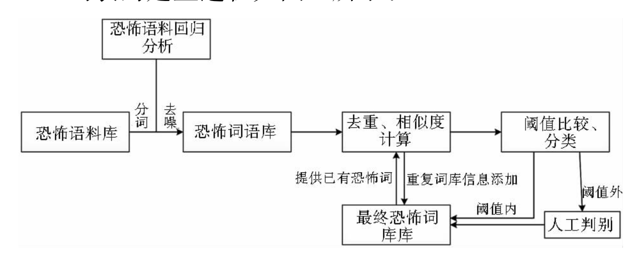


图 7 反恐词库建立流程图

实验过程如下:

(1)通过回归分析对恐怖信息词库进行分词、降噪处理。

(2)根据词库中词库进行语义相似度计算和判重,将相同词的信息添加到原有词库中,新的恐怖信息词库进行语义相似度计算。

(3)与词库中已有词库进行相关度计算,把计算后的结果与定义的(0,0.85)的阈值进行对比,相关度在(0.85,1)表示与词库中词库语义相关度高,则直接添加到词库中,如果相关度在(0,0.85)表示与词库中词相关度不是很高,则进行人工判断,将人工判断和分析后的恐怖信息词加入到恐怖信息最终的词库中。

通过对网络恐怖信息的搜集与整理发现,网络恐怖信息的产生发展和演变是存在一定规律的,恐怖信息潜藏在网络中都会按照其关键词和链接为标签式的网络状分布,所以本文提出建立恐怖信息词表,词表的

建立是与知网人物关系图相似的网络图，按照重要文本中包含的词库的个数和链接数以及被链接数，将不同的词库分为不同层次，如图 8 所示：

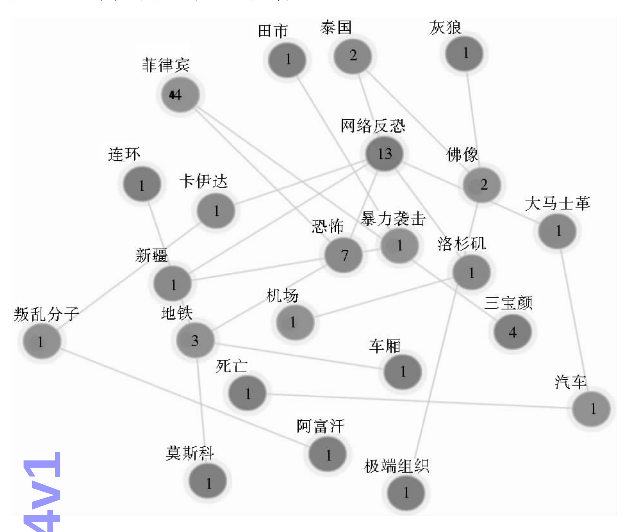


图 8　网络反恐词库关系图

图中一个圆圈代表一个关键词，圆圈中出现的数字表示该关键词对应的历史网络恐怖信息文档数，连线表示的是恐怖信息关键词之间的关系，被链接的数量反应在表 1 中。网络恐怖信息词表的建立是基于图 1 中所示网络恐怖信息数字化特征中的关联性和分散性，对网络恐怖信息的主题词、URL、本体内容建立对应关系，一方面可以为后续分析网络恐怖信息特点，针对网络恐怖信息做出网络反恐措施，另一方面也为下一次网络恐怖信息的采集提供理数据参考和线索指导，提高下一次网络恐怖信息的采集效率。

5.3　查全率与查准确率

主题网络爬虫常用的性能判断指标为查准率和查全率。查准率公式为 P = K/N，K 为抓取的与主题相关的页面数量，N 为抓取到的全部页面数。查全率又叫召回率，计算公式为：R = K/R，R 为网络中存在的所有与主题相关的页面数。为了保证 2 000 条测试数据源能够被有效利用，实验选取各类恐怖文本信息以及非恐怖文本信息中具有较大链出数量的有代表性的 URL 作为采集的初始 URL，关键词为网络反恐词库中已经建立的关键词，最终将 2 000 条包含恐怖信息的文本经过三个爬虫算法进行爬取并比较得到实验结果，见表 2。

表 2 中 PageRank 链接的相关阈值为 0.3，Δf(z)V(0.2)中 Δf(z)表示的是恐怖网页信息相似性回归中，实际涉恐网络恐怖信息 f(z)的值与标准网络恐怖信息 $f(z)_0$ 的绝对值，即 $\Delta f(z) = |f(z) - f(z)_0|$。V 表示的是恐怖关键词与涉恐 URL 回归中回

表 2　爬虫算法实验结果对比

爬虫类型	查全率	查准确率
通用爬虫	81%	61%
基于 PageRank 算法	76%	58%
回归分析算法 Δf(z)V = 0.05	65%	91%
Δf(z)V = 0.15	78%	87%
Δf(z)V = 0.2	82%	85%
Δf(z)V = 0.225	90%	71%

归预测 URL 队列的容差值，这里 V 的值恒为 0.25，V 值的选取依赖于初始 URL 的质量和词库中涉恐词库与初始 URL 对应的数量。由表中可见 Δf(z)V(0.2)的选取很重要，当其值在 0.2 左右，也就是 V = 0.25，Δf(z) = 0.8 的时候，查全率和查准确率都有较好的表现，相比于通用爬虫以及基于 PageRank 算法的爬虫，采用回归分析算法的爬虫算法能够很大程度提高同类网络恐怖信息采集的准确率，避免盲目在海量网络数据中进行采集，提高了信息采集的效率。表 2 所对应的折线图如图 9 所示：

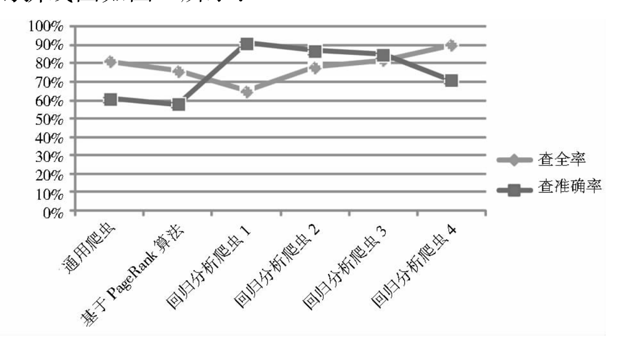


图 9　主题爬虫算法查全率与查准确率对比图

5.4　讨论与分析

从图 9 中可以看出：①本文提出的基于 PageRank 算法和 TF-IDF 算法的回归分析算法能够很好地提高对网络中具有特性恐怖信息网页的爬取效率，查准确率确实有很大的提高。②在采用回归分析爬虫的过程中，发现改变 Δf(z)V 对结果有较大的影响。由回归分析爬虫 1 可知，增大 Δf(z)能够让新采集的网络恐怖信息与旧的网络恐怖信息有较高的相似性和关联性，但是会降低网络恐怖信息采集的数量③由回归分析 2 到回归分析 3 的这段折线看出，适当增大 Δf(z)值能够在查准确率降低比较小的时候提高查全率，本文称这个时候为均衡期，一般为了避免漏网之鱼，选取 Δf(z)的值应在均衡期末期，以达到爬取所有与恐怖信息文本特性相同的文本类信息的目的。Δf(z)经过调整以后，能够充分体现到恐怖信息文本间的差异与特性，将不同类型的恐怖信息文本进行相似度比对，让主题

爬虫针对性的采集网络恐怖信息,提高网络恐怖信息文本挖掘的效率。

6 结语

随着网络信息化时代的到来,网络信息呈现爆炸式增长。恐怖分子乘机通过互联网传播恐怖组织信息或恐怖舆论信息,对国家和社会造成巨大威胁,如何从海量的开源的网络信息中搜集到与恐怖分子相关的有用信息,并将这些信息进行分析并对网络恐怖事件进行预防和制止是目前有待解决的问题。本文利用回归算法的爬虫系统能够从海量的网络信息中爬取到有价值的网络恐怖信息,为反恐工作提供情报和依据。该算法是基于已有的 PageRank 和 HITS 核心算法以及主题爬虫策略的综合运用,使不相关的算法变量,通过回归分析预测,将若干因素进行融合转化成为概率,从而找到恐怖信息相互之间的联系,设置一定大小的阈值去筛选出网络恐怖信息,达到提高信息的采集效率的目的。该算法的不足之处在于两种算法的同时使用和回归分析算法的结合使得算法过于冗杂,实际执行速度慢,对硬件的要求较高。如何对算法进行优化,提高运行效率和信息匹配精准度是我们下一步要攻克的难题。

参考文献:

[1] CÔTÉ-BOUCHER K. The diffuse border: intelligence-sharing, control and confinement along Canada's smart border [J]. Surveillance & society, 2008, 5(2):142-165.
[2] 李本先,江成俊,方锦清. 网络科学在反恐研究中面临的挑战和机遇[J]. 复杂系统与复杂性科学,2014,11(1):60-66.
[3] 李鸥. 网络反恐及对策[J]. 江西警察学院学报,2006(3):92-95.
[4] 汪勇,梅建明. 当前反恐斗争的特点、挑战及应对策略[J]. 中国人民公安大学学报(社会科学版),2016,32(1):19-23.
[5] 黄炜,余辉,李岳峰. 网络反恐知识库构建研究[J]. 情报杂志,2017,36(5):168-174.
[6] 刘炯. 网络时代暴恐音视频传播防控研究[J]. 中国人民公安大学学报(社会科学版),2015,31(1):1-9.
[7] 黄炜,余辉,李岳峰. 国内网络反恐研究的现状、问题和展望[J]. 现代图书情报技术,2016(11):1-10.
[8] CHAKRABARTI S, BERG M V D, DOM B. Focused crawling: a new approach to topic specific resource discovery[J]. Computer networks, 2000, 31(11/16):1623-1640.
[9] HEYDON A, NAJORK M. Mercator: a scalable, extensible Web crawler[J]. World Wide Web: Internet & Web information systems, 1999, 2(4):219-229.
[10] AVRAAM I, ANAGNOSTOPOULOS I. A comparison over focused Web crawling strategies[C]//Panhellenic conference on Informatics. Kastoria:IEEE Computer Society, 2011:245-249.
[11] 杨彬,康慕宁. 基于概念的权重 PageRank 改进算法[J]. 情报杂志,2006(11):70-72.
[12] 林泓,刘朋,李晶晶,龙振海. 基于概率的 PageRank 改进算法[J]. 武汉理工大学学报,2009(3):81-83.
[13] 何明,周军,李树友. 语义相似的 PageRank 改进算法[J]. 计算机工程与应用,2009(27):140-142.
[14] 王钟斐,王彪. 基于锚文本相似度的 PageRank 改进算法[J]. 计算机工程,2010(24):258-260.
[15] 王建雄. 基于特殊主题的 PageRank 改进算法[J]. 图书情报工作,2012,56(21):114-118.
[16] 王冲, 纪仙慧. 基于用户兴趣与主题相关的 PageRank 算法改进研究[J]. 计算机科学, 2016, 43(3):275-278.
[17] 王德广,周志刚,梁旭. PageRank 算法的分析及其改进[J]. 计算机工程,2010,36(22):291-293.
[18] Maestre J M, Ishii H. A PageRank based coalitional control scheme [J]. International journal of control automation & systems,2017, 15(5):1983-1990.
[19] 路永和,李焰锋. 改进 TF-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作,2013,57(3):90-95.
[20] 王景中,邱铜相. 基于 TF-IDF 改进算法的聚焦主题网络爬虫[J]. 计算机应用,2015,35(10):2901-2904,2919.
[21] 黄炜,程宝生,杨青. 基于本体的网络群体性事件主题发现研究[J]. 图书情报工作,2012,56(20):47-52,27.
[22] 李稚楹,杨武,谢治军. PageRank 算法研究综述[J]. 计算机科学,2011,38(S1):185-188.
[23] 宋聚平,王永成,尹中航,等. 对网页 PageRank 算法的改进[J]. 上海交通大学学报, 2003(3): 397-400.
[24] 朱颢东,丁温雪,杨立志,等. 微博环境下基于用户行为与主题相似度的改进 PageRank 算法[J]. 计算机工程,2017,43(5):179-184.
[25] DENG Z H, TANG S W, YANG D Q, et al. A linear text classification algorithm based on category relevance factors[C]//International conference on Asian digital libraries: digital libraries: people, knowledge, and technology. New York: Springer-Verlag, 2002:88-98.
[26] 赵小华,马建芬. 文本分类算法中词语权重计算方法的改进[J]. 电脑知识与技术,2009,5(36):10626-10628.
[27] YE Y X. New research advances in technologies of semantic Web search[J]. Computer science, 2010,1(37):1-5.
[28] 张环,刘乃文,段会川. 基于 T-Graph 算法的主题爬虫研究[J]. 计算机工程与设计,2014,35(9):3014-3017,3028.

作者贡献说明:

黄炜:提出文章思路,初步撰写内容;
张展程:撰写研究现状,负责算法实现;
朱彬:整理和分析网络恐怖信息,提供实验数据;
李岳峰:研究算法,优化论文;
陆薇:分析实验数据。

黄炜，张展程，朱彬，等.基于回归分析的网络恐怖信息主题爬虫[J].图书情报工作,2018,62(4):121 - 129.

A Network Counter-terrorism Information Crawler Based on the Regression Analysis

Huang Wei[1] Zhang Zhancheng[1] Zhu Bing[1] Li Yuefeng[1] Lu Wei[2]
[1] School of Economics and Management, Hubei University of Technology, Wuhan 430068
[2] Wuhan East Lake High-tech Development Zone Power Company, State Grid Corporation of China, Wuhan 430073

Abstract: [**Purpose/significance**] Aiming at the problems that getting the terrorist information on the network is difficult and the acquisition efficiency is low from the open source network information, a method based on the regression analysis is proposed to improve the acquisition efficiency of the network terror information by combining the advantages of the semantic relevance and the web page importance. [**Method/process**] By analyzing and comparing the characteristics of the theme crawler and combining them with the characteristics of the network terrorist information, the advantages of the PageRank algorithm and the IF-IDF algorithm for the collection of the terrorist information were found out. Combined with the regression analysis, the relevance prediction of the terrorist information was done, which reflected the process of the information collection. [**Result/conclusion**] Both the quantity and quality of the collection of the network terrorist information should be taken into consideration. Based on the traditional common network crawler algorithm, this paper proposes a crawler optimization algorithm pertinent to the network terrorist information collection, which improves the collection efficiency.

Keywords: theme crawler regression analysis network anti-terrorism semantic similarity